Hardness of discrepancy computation and epsilon-net verification in high dimension

Panos Giannopoulos[*]   Christian Knauer[†‡]   Magnus Wahlström[§¶]

Daniel Werner[‖**]

January 16, 2013

Abstract

Discrepancy measures how uniformly distributed a point set is with respect to a given set of ranges. There are two notions of discrepancy, namely *continuous discrepancy* and *combinatorial discrepancy*. Depending on the ranges, several possible variants arise, for example star discrepancy, box discrepancy, and discrepancy of half-spaces. In this paper, we investigate the hardness of these problems with respect to the dimension d of the underlying space.

All these problems are solvable in time $n^{O(d)}$, but such a time dependency quickly becomes intractable for high-dimensional data. Thus it is interesting to ask whether the dependency on d can be moderated.

We answer this question negatively by proving that the canonical decision problems are W[1]-hard with respect to the dimension. This is done via a parameterized reduction from the CLIQUE problem. As the parameter stays linear in the input parameter, the results moreover imply that these problems require $n^{\Omega(d)}$ time, unless 3-SAT can be solved in $2^{o(n)}$ time. Further, we derive that testing whether a given set is an ε-net with respect to half-spaces takes $n^{\Omega(d)}$ time under the same assumption. As intermediate results, we discover the W[1]-hardness of other well known problems,

such as determining the largest empty star inside the unit cube. For this, we show that it is even hard to approximate within a factor of 2^n.

1   Introduction

Geometric discrepancy has significant applications in several areas, including optimization, statistics, combinatorics, and computer graphics. See for example the textbooks by [Mat10], [Cha00], and [DT97]. In particular, the *star discrepancy* of a point set is important in multi-variate numerical integration, where the error of a quasi-Monte Carlo integration is bounded as a function of the star discrepancy of the point set used in the integration (by the Koksma–Hlawka inequality, see [Nie92]). In addition, the difficulty of computing the star discrepancy can be an obstacle to evaluating different methods for creating low-discrepancy point sets, see, for example, [DGW10].

Unfortunately, computing the star discrepancy of a point set using any known method is computationally intensive; given a set P of n points in d dimensions, every known method for getting even a constant-factor approximation of its star discrepancy has a running time of $n^{\Theta(d)}$ (see below). As many applications, for example in financial mathematics, require integration of functions in tens or even hundreds of dimensions, this quickly becomes infeasible.

The main question we ask here is whether this dependency on d is necessary. Specifically, we ask whether the decision version for star discrepancy (and other related problems) can be solved in $O(f(d)n^c)$ time, for some computable function f and some constant c independent of d, i.e, whether it is fixed-parameter tractable with respect to d. Note that NP-hardness for a problem does not exclude such a possibility. Proving that a problem is W[1]-hard with respect to d implies that such an algorithm is not possible, under standard complexity theoretic assumptions.

1.1   Parameterized Complexity.

We review some basic definitions of parameterized complexity theory; see, for example, one of the textbooks by [DF99], [FG06], and [Nie06] for an introduction. A problem with input size n and a positive integer parameter k is *fixed-parameter tractable* (fpt for short) if it can be solved by an algorithm that runs in $O(f(k) \cdot n^c)$ time, where f is a computable function depending only on k, and c is a constant independent of k; such an algorithm is (informally) said to run in fpt-time. The class of all fixed-parameter tractable problems is denoted by FPT.

An infinite hierarchy of classes, the W-hierarchy, has been introduced for establishing fixed-parameter intractability. Its first level, W[1], can be thought of as the parameterized analog of NP: a parameterized problem that is hard for W[1] is not in FPT unless

FPT=W[1], which is considered highly unlikely under standard complexity theoretic assumptions. Hardness is sought via an *fpt-reduction*, i.e., an fpt-time many-one reduction from a problem Π, parameterized with k, to a problem Π', parameterized with k', such that $k' \leq g(k)$ for some computable function g.

1.2 Discrepancy and epsilon-nets

In this section, we define the basic notion of discrepancy. Let X be set and $\mathcal{R}$ be a set of subsets of X, both not necessarily finite. A tuple $(X, \mathcal{R})$ is called a *range space*.

If the range space arises from point sets and geometric objects, such as half-spaces or hyperrectangles (boxes), we call it a *geometric range space*. Often, as in our case, the ranges are given implicitly. As an example, for a point set $P \subset \mathbb{R}^d$, we define

$$\mathcal{H}_P := \{H \cap P \mid H \text{ is a half-space}\} .$$

This is the range space induced by all half-spaces in $\mathbb{R}^d$. Observe that if P is finite, even though there are infinitely many half-spaces, the size of $\mathcal{R}$ is at most $2^{|P|}$.

We will now define two different notions of discrepancy, namely the *continuous discrepancy* and the *combinatorial discrepancy*.

1.2.1 Continuous Discrepancy

This concept relates the volume (i.e., its Lebesgue-measure) of a point set to its discrete measure, i.e., to the fraction of points it contains. To simplify matters, we restrict ourselves to point sets in the d-dimensional unit cube.

The intuition is the following: a range space should have high discrepancy either if there is a range with small volume that contains a large fraction of points, or if there is a range with large volume that contains a small fraction of points. In that sense, it measures how good a finite set of points approximates the uniform distribution.

Definition 1. *Let $P \subseteq \mathbb{R}^d$ be a set of points and $\mathcal{R}$ be a set of subsets of $[0, 1]^d$. We define the* continuous discrepancy *of P with respect to $\mathcal{R}$ as*

$$\bar{D}_{\mathcal{R}}(P) := \max_{R \in \mathcal{R}} \left| vol(R) - \frac{|R \cap P|}{|P|} \right| .$$

Figure 1 shows two point sets in the plane with respect to axis-parallel boxes. The point set in Figure 1(a) shows a point set with low discrepancy. In 1(b), the discrepancy is attained for a box of high volume and few points inside. In 1(c), it is attained for a small volume box with many points inside.

1.2.2 Combinatorial Discrepancy

The *combinatorial discrepancy*, sometimes called red-blue-discrepancy, is a slightly different notion. Here, we are given a set of points P, colored red or blue, and a set of ranges. Such

(a) low discrepancy set   (b) large box, few points   (c) small box, many points
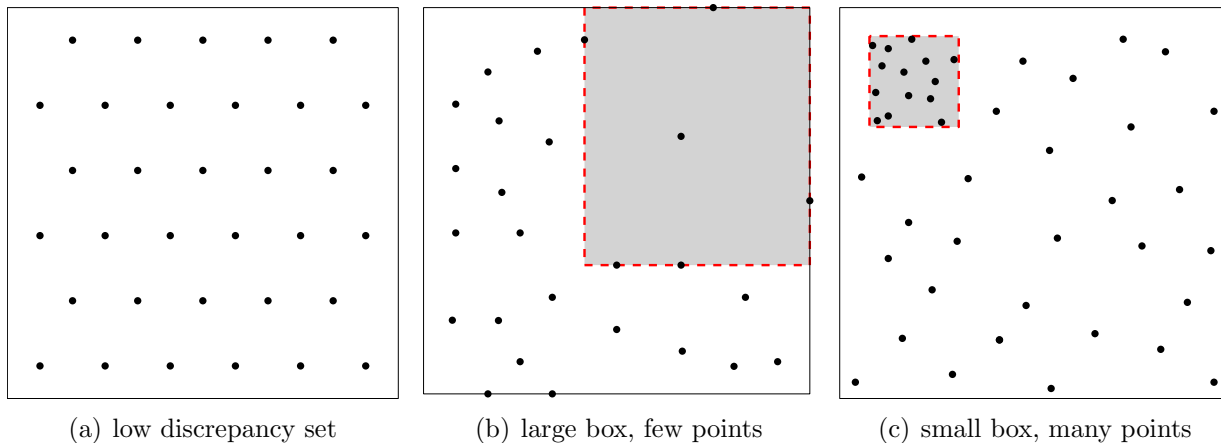
Figure 1: Continuous Discrepancy

a set is said to have high discrepancy, if there is a range where the difference between red and blue points is high.

Definition 2. *Let* $P = P_r \uplus P_b$ *be a set of points in* $\mathbb{R}^d$, *and let* $\mathcal{R}$ *be a set of subsets of* P. *We define the* combinatorial discrepancy *of* P *with respect to* $\mathcal{R}$ *as*

$$\dot{D}_{\mathcal{R}} (P_r, P_b) := \max_{R \in \mathcal{R}} ||R \cap P_r| - |R \cap P_b||.$$

1.3 Epsilon-nets

A theory closely related to discrepancy is that of ε-nets. We will give some basic terminology and afterwards discuss the relation of our results to ε-net problems.

For a range space $(P, \mathcal{R})$, a set $S \subseteq P$ is called an ε-net if for all $R \in \mathcal{R}$ with $|R| \geq \varepsilon|P|$ it holds that $R \cap S \neq \emptyset$. This means that S intersects all large sets in $\mathcal{R}$, namely those that contain at least an ε-fraction of the points.

For applications, one is of course interested in nets that are small. For general range spaces, one can not expect such a behavior: If $\mathcal{R}$ is the power-set of P, any ε-net must be of size at least $n - \varepsilon n + 1$.

Surprisingly, there are many range spaces where the size of a net does not depend on the value of n. The *Vapnik-Chervonenkis dimension* has proved as a useful tool for studying these range spaces. Let $(P, \mathcal{R})$ be a range space. For a set $Q \subset P$, let $\text{pr}_Q(P) := \{R \cap Q \mid R \in \mathcal{R}\}$ denote the projection onto Q. We say that a set Q is *shattered* by $\mathcal{R}$, if $\text{pr}_Q(P) = \mathcal{P}(Q)$, i.e., if all subsets of Q appear in the projection of P onto Q. Now the VC-dimension δ of a range space is the size of the largest set $Q \subset P$ that is shattered by $\mathcal{R}$.

[HW86] proved, based on a work by [VC71], that range spaces of finite VC-dimension δ admit ε-nets of size $O\left(\frac{\delta}{\varepsilon} \log \frac{\delta}{\varepsilon}\right)$. Their result even proves something a lot stronger, namely that a random sample of that size will be an ε-net with high probability.

The theory of ε-nets has been a hot topic recently, most notably because of the results by [Alo11] and [PT10].

1.4 Our results

We study the decision versions of several problems related to the above definitions. In Section 2, we consider the following problem.

Definition 3. *(d-*RED-BLUE-DISCREPANCY*) Let* $P = P_r \uplus P_b$ *be set of points in* $\mathbb{R}^d$, *and* $k \in \mathbb{N}$. *Decide whether* $\dot{D}_{\mathcal{B}}(P_r, P_b) \geq k$, *where* $\mathcal{B}$ *is the set of all axis-parallel boxes inside the unit cube.*

In particular, we show the following.

Theorem 1. *The problem d-*RED-BLUE-DISCREPANCY *is W[1]-hard with respect to the dimension and NP-hard.*

This result will be easily derived by first showing the W[1]-hardness of another well known problem, called BICHROMATIC-RECTANGLE, where we have to find a box that contains as many blue points as possible, but no red points.

Subsequently, we will investigate the case of continuous discrepancy. Thereto, let $\mathcal{B}_0$ be the set of axis-parallel boxes inside the unit cube containing the origin. Such a box is called a *star*, or *subinterval*. In Section 4 and Section 6, we will consider the following problem.

Definition 4. *(d-*STAR-DISCREPANCY*) Let* P *be a set of points in* $\mathbb{R}^d$ *and* V *be a rational number. Decide whether* $\bar{D}_{\mathcal{B}_0}(P) \geq V$.

The problem where the range space is the set of *all* axis-parallel boxes inside the unit cube $\mathcal{B}$ is defined analogously and will be called *d-*BOX-DISCREPANCY. In Sections 4 and 6, we show the following. (The NP-hardness of star discrepancy was shown by [GSW09].)

Theorem 2. *The problems d-*STAR-DISCREPANCY *and d-*BOX-DISCREPANCY *are W[1]-hard with respect to the dimension and NP-hard.*

In order to prove these two theorems, we consider two related problems, which have also been studied in the past.

Definition 5. *(d-*MAXIMUM-EMPTY-STAR*) Let* P *be a set of points in* $\mathbb{R}^d$ *and* V *be a rational number. Decide whether there is a box of volume* V *inside the unit cube that contains the origin but no points from* P.

Analogously, we define the problem *d-*MAXIMUM-EMPTY-BOX. We establish the following results.

Theorem 3. *The problems d-*MAXIMUM-EMPTY-STAR *and d-*MAXIMUM-EMPTY-BOX *are W[1]-hard with respect to the dimension.*

For the d-Maximum-Empty-Star problem we immediately get a result that is a lot stronger:

Theorem 4. *The problem d-Maximum-Empty-Star cannot be approximated in fpt time within a factor of $2^{|P|}$, unless $FPT = W[1]$.*

Afterwards, we sketch hardness results for some other range spaces, such as boxes or half-spaces.

Finally, we build the connection to the theory of ε-nets. It is known that for many sets, a small random sample will be an ε-net with high probability. This leads to the following decision version.

Definition 6. *(d-Net-Verification) Let P be a set of points in $\mathbb{R}^d$, S be a subset of P, and an $\varepsilon > 0$. Decide whether S is an ε-net for P with respect to half-spaces.*

Our main theorem in this section shows that this question cannot be answered efficiently.

Theorem 5. *The problem d-Net-Verification is co-NP-hard and co-W[1]-hard with respect to the dimension.*

In all our reductions, the parameter d is kept linear in the input parameter. Using a result by [CCF+05], we can derive something stronger.

Corollary 1. *All of the above problem cannot be solved in time $n^{o(d)}$, unless the Exponential Time Hypothesis fails, i.e., unless 3-Sat can be solved in $2^{o(n)}$ time.*

These results are obtained by fpt-reductions from the W[1]-complete k-Clique problem in general graphs, see [DF99], based on the general framework by [CGK+11, CGKR08].

1.5 Related work.

When the dimension is part of the input, the Bichromatic-Rectangle problem was shown to be NP-hard by [EHL+02]; in the same paper an $O(n^{2d+1})$-time algorithm was given. [BK09] gave an algorithm that runs in $O(k \log^{d-2} n)$ time, where k is the number of feasible boxes that are not properly contained in any feasible box, and showed that k can be $\Theta(n^d)$ in the worst case. [AHP08] gave an $(1 - \varepsilon)$-approximation algorithm that runs in $O(n^{\lceil d/2 \rceil}(\varepsilon^{-2} \log n)^{\lceil d/2+1 \rceil})$ time.

The Star-Discrepancy problem has been shown to be NP-hard by [GSW09]. An exact algorithm that runs in $O(n^{1+d/2})$ time was given by [DEM96]. [Thi01] has given an approximation algorithm that achieves additive error and runs in fpt-time with respect to the error and the dimension. However, as [Gne08] noted, by setting the error tolerance to the same order as the discrepancy of an optimal point set, so that a constant-factor approximation is achieved, the running time of any algorithm following Thiémard's approach becomes $n^{O(d)}$. As for the Box-Discrepancy, no hardness results where known so far.

The MAXIMUM-EMPTY-BOX problem has been studied extensively in the planar case, see for example [AS87] and references therein. When the dimension is part of the input, the problem has only recently been shown to be NP-hard by [BK10] and the fastest exact algorithm runs in time $O(n^d \log^{d-2} n)$ [BK10]. Also recently, [DJ09] gave an $O((8ed\varepsilon^{-2})^d \cdot n \log^d n)$-time $(1-\varepsilon)$-approximation algorithm for this problem. Note that, since $(\log n)^k < n + f(k)$ for some $f(k)$, this counts as fpt time in parameters $1/\epsilon$ and d, in contrast to our results for MAXIMUM-EMPTY-STAR. The NP-hardness of the MAXIMUM-EMPTY-STAR problem was shown by [GSW09].

2 Red-Blue Discrepancy and the Bichromatic Rectangle Problem

In order to show the hardness of RED-BLUE-DISCREPANCY, we will first consider the BICHROMATIC-RECTANGLE problem. The parameterized decision problem is defined as follows:

Definition 7. *(k-d-BICHROMATIC-RECTANGLE) Let P_r be a set of red points and a set P_b be a set of blue points in $\mathbb{R}^d$, and $k \in \mathbb{N}$. Decide whether there is an axis-parallel box such that $H \cap P_r = \emptyset$ and $|H \cap P_b| \geq k$?*

A box that does not contain any point from P_r will be called *feasible*. For a given set of points $P = P_r \uplus P_b$, let

$$E_{\mathcal{B}}(P_r, P_b) = \max_{B \in \mathcal{B}, B \cap P_r = \emptyset} |B \cap P_b|$$

denote the size of an optimal solution. Recall that $\mathcal{B}$ is the set of all axis-parallel boxes inside the unit cube.

2.1 The idea.

In order to show that the k-d-BICHROMATIC-RECTANGLE problem is $W[1]$-hard, we will give a reduction from the k-CLIQUE problem. For a given simple graph $G = ([n], E)$ we will construct sets $P_r = P_r(G, k)$ and $P_b = P_b(G, k)$ in $\mathbb{R}^{2k}$ such that G has a clique of size k if and only if $E(P_r, P_b) = k + 1$.

In addition to the (blue) origin 0, we will put blue and red points into k pairwise orthogonal two-dimensional planes. These points will be used to encode the vertices of G. Additional red points will then be used to encode the edge-set of G.

Each plane will contain n blue points, corresponding to the vertices of the graph, and $n-1$ red points. The red points are placed such that no feasible box can contain more than one blue point from a single plane. Thus, at most k of these blue points can be contained in any feasible box. We will then ensure that such a box can only contain points x and y from two different planes if the corresponding vertices are connected in G. This is done by

putting red points into the product of the respective planes (which is a four-dimensional subspace).

This construction will ensure that any feasible box containing $k+1$ blue points corresponds to a k-clique in G, and vice versa.

2.2 Preparations.

For $1 \leq i \leq k$, we define the two-dimensional subspace

$$\mathbb{R}_i^2 = \{(x_1, y_1, \ldots, x_k, y_k) \mid x_j = y_j = 0, j \neq i\} \subseteq \mathbb{R}^{2k}.$$

For $1 \leq i < j \leq k$, we set $\mathbb{R}_{ij}^4$ to be the product of $\mathbb{R}_i^2$ and $\mathbb{R}_j^2$, i.e., $\mathbb{R}_{ij}^4 = \mathbb{R}_i^2 \times \mathbb{R}_j^2$. For $p \in \mathbb{R}_i^2$ and $q \in \mathbb{R}_j^2$, observe that the unique point in $\mathbb{R}_{ij}^4$ that (orthogonally) projects to p (into $\mathbb{R}_i^2$) and to q (into $\mathbb{R}_j^2$) is $p + q$.

2.3 The scaffold construction.

Let $\varepsilon = 1/4$. For a vertex $1 \leq v \leq n$, we define the point

$$b_i(v) = (v, n + 1 - v) \in \mathbb{R}_i^2.$$

Then, let

$$(P_b)_i^{\text{scaffold}} = \{b_i(1), \ldots, b_i(n)\} \subseteq \mathbb{R}_i^2$$

be the set of all points in the i-th plane. Choosing a (rectangle containing) point $b_i(v)$ will correspond to choosing vertex v from G. Let

$$P_b^{\text{scaffold}} = \biguplus_{1 \leq i \leq k} (P_b)_i^{\text{scaffold}}$$

be the set of all these blue points.

As we want the feasible boxes to contain at most one point from each $\mathbb{R}_i^2$, we add a set of red points as follows: For $1 \leq v \leq n - 1$, we define $r_i(v) = (v + 1/2, n + 1 - (v + 1/2))$ and set

$$(P_r)_i^{\text{scaffold}} = \{r_i(1), \ldots, r_i(n-1)\} \subseteq \mathbb{R}_i^2.$$

Finally, we define

$$P_r^{\text{scaffold}} = \biguplus_{1 \leq i \leq k} (P_r)_i^{\text{scaffold}}$$

to be the set of all red scaffolding points. See Figure 2 for an example of the scaffold construction. Observe that an feasible box B can contain at most one blue point from each $(P_b)_i^{\text{scaffold}}$.
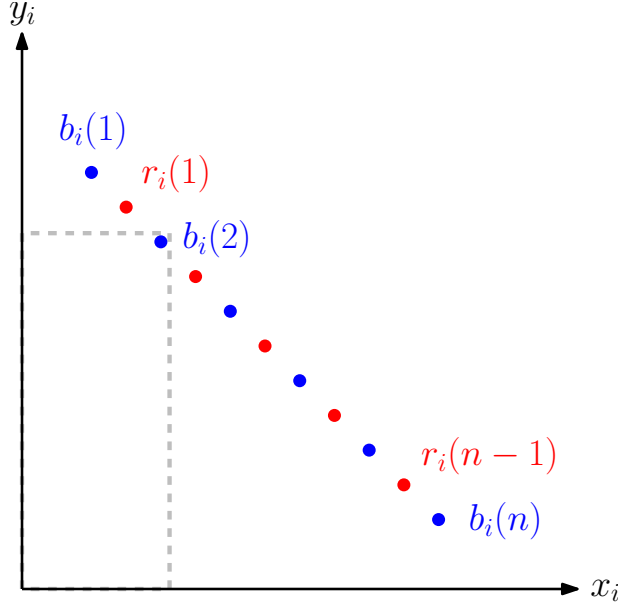
8

Figure 2: The scaffold construction with vertex 2 selected.

2.4 Encoding edges.

In order to encode the edges of the graph, we will place several red points between pairs of $\mathbb{R}_i^2$s. This will forbid certain pairs of blue points to be selected at the same time, namely the ones that correspond to vertices not being connected in G.

Thereto, for $1 \leq i \leq j \leq k$ and vertices $1 \leq u, v \leq n$, we define the point

$$r_{ij}^{\mathrm{kill}}(uv) = b_i(u) + b_j(v) \in \mathbb{R}_{ij}^4.$$

The crucial property of such a point is that it is inside a box (containing the origin) if and only if both $b_i(u)$ and $b_j(v)$ are inside this box.

The set of all killing points in $\mathbb{R}_{ij}^4$ is then

$$(P_r)_{ij}^{\mathrm{E}} = \{r_{ij}^{\mathrm{kill}}(uv), r_{ij}^{\mathrm{kill}}(vu) \mid uv \notin E\}.$$

As the graph is simple (i.e., contains no loops), so all points of the form $r_{ij}(uu)$ are also added. Finally, we set

$$P_r^{\mathrm{E}} = \biguplus_{1 \leq i \neq j \leq k} (P_r)_{ij}^{\mathrm{E}}$$

to be the set of all killing points. See Figure 3 for an example where $uv \notin E$.

2.5 The overall construction.

For $G = ([n], E)$ and $k > 0$ we construct point sets $P_r(G, k), P_b(G, k)$ in $\mathbb{R}^{2k}$ as follows:
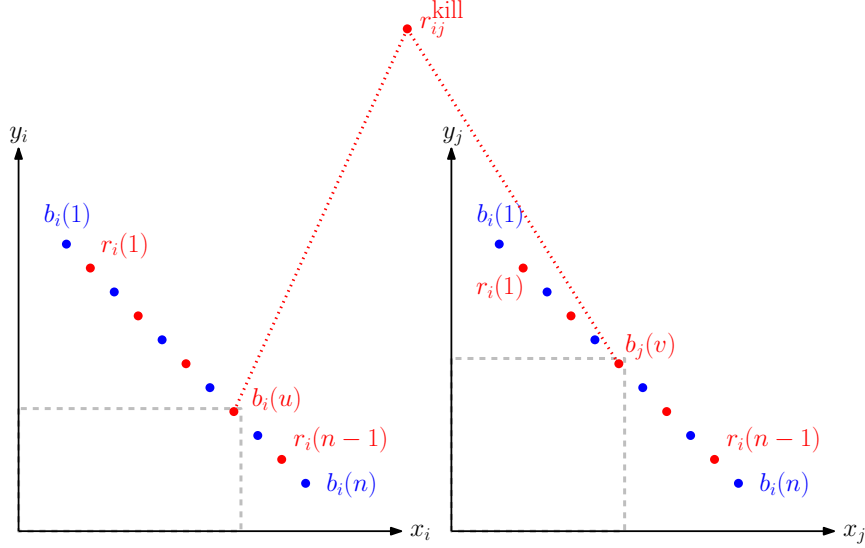
- $P_r(G, k) = P_r^{\mathrm{scaffold}} \cup P_r^{\mathrm{E}}$

Figure 3: $b_i(u)$ is the projection of $r_{ij}^{\mathrm{kill}}(uv)$ to $\mathbb{R}_i^2$ and $b_j(v)$ is the projection of $r_{ij}^{\mathrm{kill}}(uv)$ to $\mathbb{R}_j^2$.

- $P_b(G, k) = \{0\} \cup P_b^{\mathrm{scaffold}}$

The size of the point set is $O(k^2 n^2)$ and the coordinates of the points can be encoded by $O(\log kn)$ many bits. Clearly the construction can be performed in time polynomial in both k and n.

Lemma 1. *G has a k-clique if and only if $E_{\mathcal{B}}(P_r, P_b) = k + 1$.*

Proof. First, observe that any feasible box B can contain at most $k + 1$ points, as $|B \cap (P_b)_i^{\mathrm{scaffold}}| \leq 1$, for $1 \leq i \leq k$. Additionally, 0 can be in B.

Let $v_1, \ldots, v_k$ be a clique of size k. We choose a (closed) box B with upper right corner $b_i(v_i)$ in $\mathbb{R}_i^2$.

B contains exactly one point from each of the $\mathbb{R}_i^2$, and also the origin, making it a total of $k+1$ points. We show that B is feasible. First, by definition B contains no point of P_r^{scaffold}. Further, assume that B contains a point of P_r^E, say $r_{ij}^{\mathrm{kill}}(uv) = b_i(u) + b_j(v) \in (P_r)_{ij}^E$. Then B contains both $b_i(u)$ and $b_j(v)$. But this means $uv \notin E$, for otherwise the point $r_{ij}^{\mathrm{kill}}(uv)$ would not have been added; a contradiction.

Now assume that there is no clique of size k. Let B be any box containing $k+1$ points. We show that B is infeasible. If B contains a red point from one of the $\mathbb{R}_i^2$'s, we are done. Otherwise, besides the origin it can contain at most one blue point from each $\mathbb{R}_i^2$. Let $u = v_i$ and $v = v_j$ be two vertices corresponding to blue points contained in B that are not connected in G. As there is no k-clique, such a pair must exist. Then B also contains the red point $r_{ij}^{\mathrm{kill}}(uv)$. Thus, B is infeasible. $\qquad\square \qquad\qquad\qquad \square$

Theorem 6. *The k-d-Bichromatic-Rectangle problem is $W[1]$-hard when parameterized with both the dimension d and the size of the solution k.*

As noted by [CT97] an optimization problem that is $W[1]$-hard when parameterized by the size of the solution is unlikely to have an approximation scheme that runs in $f(\varepsilon)n^c$, i.e, an efficient polynomial-time approximation scheme (EPTAS). In our case, since the problem is hard with respect to both the dimension and the size of the solution, this implies the following:

Corollary 2. *The (optimization version of the) k-d-BICHROMATIC-RECTANGLE problem does not admit an approximation scheme that runs in $O(f(1/\varepsilon, d) \cdot \mathrm{poly}(n, d))$ time, unless $W[1] = FPT$.*

2.6 Adaption to Red-Blue Discrepancy

In order to adapt this proof to the RED-BLUE-DISCREPANCY problem, we have to modify the set in such a way that a clique corresponds to a set with high discrepancy. Thereto, let N be the total number of points in the construction. We replace the blue point at the origin by N copies. Now the value

$$\dot{D}_{\mathcal{R}}\left(P_r, P_b\right) := \max_{R \in \mathcal{R}} ||R \cap P_r| - |R \cap P_b||$$

is maximized for a box with many blue points inside (as it is at least N). Observing that, for each $\mathbb{R}_i^2$, the difference between blue and red points is at most one, we can follow the above reasoning. This means there is a box with $N + k$ blue points (and no red points) inside, if and only if G has a k-clique. This proves Theorem 1.

3 The Maximum Empty Star Problem

We now turn to the continuous version of the problems. In this section, we consider the problem where we have to find an empty axis-parallel box inside the unit-cube of maximal volume that contains the origin, namely MAXIMUM-EMPTY-STAR. Besides showing the $W[1]$-hardness, our construction yields that it is even $W[1]$-hard to approximate this problem by a factor of $\left(\frac{1}{2}\right)^{|R|}$.

The proof uses ideas similar to the discrete case in Section 2. As above, because the origin has to be included in the boxes, the planes will be considered separately. In this construction, the analog of a rectangle containing a blue point from one of the $\mathbb{R}_i^2$ is now a rectangle that is "large" (of size C for some $0 < C < 1$ to be determined later). In each plane, there will be n large rectangles to choose from, corresponding to the n vertices of G. It will only be possible to choose large rectangles from two different planes, if the corresponding vertices are connected in G. This yields a one-to-one correspondence between "large" empty boxes and cliques of size k.

3.1 The Construction.

We will proceed as follows: First, we determine where the upper right corners of the n large rectangles have to be. From this, we will determine the blocking points (which are

the analog of the red points above) that are needed for this.

Let $\mu > 1$ be a parameter to be specified later. One possibility to determine the upper right corners of the rectangles, each having area $C = \frac{1}{\mu^{n-1}}$ in one $\mathbb{R}_i^2$, is as follows:

$$c_i(u) = \left(C\mu^{u-1}, \frac{1}{\mu^{u-1}} \right), 1 \leq u \leq n.$$

We now place points such that any maximal empty (open) rectangle, i.e., a rectangle supported by two points, has its upper right corner at one $c_i(u)$. This can be realized by the following blocking points:

$$p_i(u) = \left(C\mu^{u-1}, \frac{1}{\mu^u} \right), 0 \leq u \leq n.$$

We set $P_i^{\text{scaffold}} = \{p_i(u) \mid 0 \leq u \leq n\}$ and $P^{\text{scaffold}} = \biguplus_{1 \leq i \leq k} P_i^{\text{scaffold}}$.

Thus, in each $\mathbb{R}_i^2$, we have n choices for the upper right corner of the rectangles: the points $c_i(u)$, $1 \leq u \leq n$. If a rectangle has its upper right point somewhere else on $(x, C/x)$ or above, it contains a point from P^{scaffold}, and any other feasible rectangle has smaller size.

Choosing a large rectangle in each of the $\mathbb{R}_i^2$ gives us an empty rectangle of total volume C^k. See Figure 4 for an example.

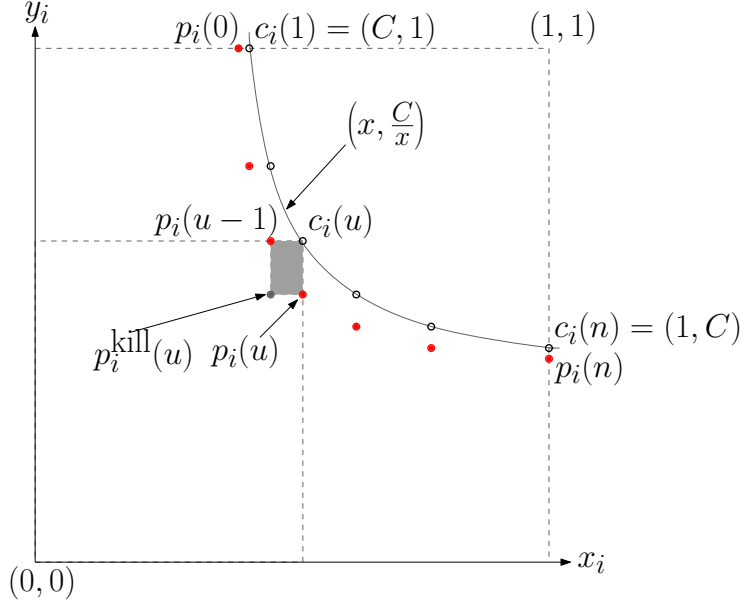
Figure 4: The plane $\mathbb{R}_i^2$. A rectangle selecting vertex u is indicated.

3.2 Encoding the edges.

As above, if the vertices corresponding to two different large rectangles in the planes $\mathbb{R}_i^2$ and $\mathbb{R}_j^2$ are not connected, we will add a point in the product $\mathbb{R}_{ij}^4$ that forbids these two

rectangles to be chosen at the same time. To this end, we let

$$p_i^{\text{kill}}(u) = (C\mu^{u-2}, 1/\mu^u)$$

(these points are themselves not added to the set P) and then define $p_{ij}^{\text{kill}}(uv) = p_i^{\text{kill}}(u) + p_j^{\text{kill}}(v)$. Recall that this also includes all points of the form $p_{ij}(uu)$. Then we set

$$P^{\text{E}} = \{p_{ij}^{\text{kill}}(uv) \mid i \neq j, uv \notin E\},$$

and

$$P = P^{\text{E}} \cup P^{\text{scaffold}}.$$

The size of P is $O(n^2 k^2)$. If we set $\mu = 2$, all coordinates have size polynomial in the size of the input. We also let $V = C^k = 1/\mu^{k(n-1)}$. Clearly the construction can be performed in time polynomial in k and n.

Now we come to prove the correctness of the construction. Let $F_i(u)$ be the rectangle with corners $p_i(u-1)$, $c_i(u)$, $p_i(u)$, $p_i^{\text{kill}}(u)$, as indicated in Figure 4.

Lemma 2. *Any feasible rectangle in $\mathbb{R}_i^2$ that does not intersect any region $F_i(u)$, $1 \leq u \leq n$, has size at most C/μ.*

Proof. Such a rectangle has its upper right point below the graph going through the points $p_i(u)$, $1 \leq u \leq n$, which is $(x, \frac{C}{x\mu})$. $\square$ $\square$

We use this to prove the main Lemma, the continuous analog of Lemma 1:

Lemma 3. *G has a k-clique if and only if there is an empty star of size $V = C^k$. Further, if G does not have a k-clique, the largest empty star has volume C^k/μ.*

Proof. Let $v_1, \ldots v_k$ be a clique in G. In each $\mathbb{R}_i^2$, $1 \leq i \leq k$, choose the rectangle with upper right corner $(C \cdot \mu^{v_i-1}, \frac{1}{\mu^{v_i-1}})$. Then the box is the product of these rectangles and it has volume C^k. By definition, it does not contain any point from one of the $\mathbb{R}_i^2$. If it would contain a point $p_{ij}(uv) \in \mathbb{R}_{ij}$, then the projection of $p_{ij}(uv)$ onto both $\mathbb{R}_i^2$ and $\mathbb{R}_j^2$ would be contained in the corresponding rectangles. But this means that $uv \notin E$, a contradiction.

If there is no k-clique, any selection of k large rectangles that do not contain a point in $\mathbb{R}_i^2$ would contain a point in one of the $\mathbb{R}_{ij}^4$, as among any k vertices there are at least two that are not connected. Thus, in order to avoid all points, by Lemma 2 in at least one $\mathbb{R}_i^2$ we cannot select a large rectangle intersecting any of the $F_i(u)$. Thus, the total volume can be at most $C^{k-1} \cdot \frac{C}{\mu} = C^k/\mu$. $\square$ $\square$

Thus, we have shown the first part of Theorem 3.

3.3 An inapproximability result.

Now it easily follows that the problem is even hard to approximate: the μ chosen above can be made very large. By Lemma 3, the ratio between a large box for a point set constructed from a positive instance and one constructed from a negative instance is at least μ. Since we can choose $\mu = 2^{O(|R|)}$ (this takes only polynomially many bits), Theorem 4 follows.

4 The Star Discrepancy Problem

In this section we show that computing the star discrepancy of a point set inside the unit cube is $W[1]$-hard.

There are two reasons why the previous reduction does not give us the hardness-result for this problem right away:

- First, the maximum discrepancy can be attained by either a large box with few points inside or by *a small box with many points inside.* For example, in our construction from Section 3, large point sets lie in a box with (affine) dimension $d - 1$ and thus a volume of 0.

- Second, even if the maximum is attained for a large box, it might still contain *some* points, in which case our construction would fail.

However, we can get rid of both problems by simply choosing the right value for μ, and thus, C. Recall that these values determined the size of the largest empty box. C is exactly the area of a maximum empty rectangle in each $\mathbb{R}_i^2$.

For a graph G, let N be the total number of points in our construction from the previous section. Recall that $N \in O(k^2 n^2)$. Any box (containing the origin) that contains all points has volume of 1, as there are points with $x_i = 1$ and $y_i = 1$ for all $1 \leq i \leq k$. This leads to the following observation.

Corollary 3. *For all boxes B containing the origin we have*

$$\frac{|B \cap P|}{|P|} - vol(B) \leq \frac{N-1}{N}.$$

That means that the fraction of points in any box can be bigger by at most $(N-1)/N$ than its volume.

Thus, our construction from Section 3 works if we can ensure that the largest empty box in a positive instance has volume more than $(N-1)/N$, that means

$$C^k = \left(\frac{1}{\mu^{n-1}}\right)^k > \left(\frac{N-1}{N}\right).$$

Then the discrepancy is attained for an empty box: Enlarging the box can increase the volume by at most $1 - C^k$. But as $C^k > \frac{N-1}{N} = (1 - \frac{1}{N})$, we have that $(1 - C^k) < 1/N$. Thus, picking such an extra point cannot increase the discrepancy. This means that we can choose μ such that

$$1 < \mu < \left(\frac{N}{N-1}\right)^{\frac{1}{k(n-1)}}.$$

To make sure that μ requires only polynomially (in k and n) many bits, observe the following.

Lemma 4. *For $\mu = 1 + \frac{1}{t}$ with $t = 2knN$, it holds that $\mu^{k(n-1)} < \frac{N}{N-1}$.*

Proof. Observe that $\frac{N}{N-1} = \sum_{i=0}^{\infty} \left(\frac{1}{N}\right)^i$. Then

$$\mu^{k(n-1)} < \mu^{kn} = \left(1 + \frac{1}{t}\right)^{kn} = \sum_{i=0}^{kn} \binom{kn}{i} \frac{1}{t^i} \le \sum_{i=0}^{kn} (kn)^i \frac{1}{t^i} \le \sum_{i=0}^{\infty} \left(\frac{kn}{t}\right)^i.$$

Thus,

$$\mu^{k(n-1)} = \left(1 + \frac{1}{t}\right)^{k(n-1)} < \sum_{i=0}^{\infty} \left(\frac{1}{2N}\right)^i < \sum_{i=0}^{\infty} \left(\frac{1}{N}\right)^i = \frac{N}{N-1}.$$

$\square$ $\square$

Constructing the set P with this value of μ, we immediately get:

Lemma 5. *G has a clique of size k, if and only if $\bar{D}_{\mathcal{B}_0}(P) = C^k$.*

Proof. By the previous remarks, the maximum is attained for a *large empty* box. Then, the proof follows from Lemma 3. $\square$ $\square$

This proves the first part of Theorem 2.

5 Largest Empty Box Problem

In the two upcoming sections, we will consider the analogous problems for the case when the origin does not have to be contained in the boxes. That means our range space is $\mathcal{B}$, the set of all axis-parallel boxes inside the unit cube, as defined in Section 1.2. We start with the case where we have to find a large empty box inside the unit cube, namely d-Maximum-Empty-Box.

This problem is quite different in the sense that, as the box does not have to contain the origin, now the $\mathbb{R}_i^2$ cannot be considered separately any more. This kills our construction from the previous section: the box $(0,1)^{2k}$ does not contain any points from P but has volume 1.

The plan is to reestablish this dependence, so that we can use the same reasoning as above. This can be done by a simple trick, which we call *lifting*: From a graph G, we first construct the set P as in Section 3 with the constant $C^k = 2/3$. Then, define the function lift $\colon \mathbb{R}^{2k} \to \mathbb{R}^{2k}$ as follows:

$$\text{lift}(x_1, \ldots, x_{2k}) = (x_1', \ldots, x_{2k}') \text{ with } x_i' = \begin{cases} x_i & \text{if } x_i \neq 0 \\ x_i' = 1/2 & \text{otherwise.} \end{cases}$$

Now we apply the function lift to all points in the set P. For the lifted point x, we call the $\mathbb{R}_i^2$ that the point was lifted from the *corresponding* $\mathbb{R}_i^2$. This gives the following:

Lemma 6. *Any box $B \in \mathcal{B}$ having volume at least $2/3$ contains a point x if and only if the projection onto the corresponding $\mathbb{R}_i^2$ contains the projection of x.*

Proof. As the box has volume at least $2/3$, each of its projections onto any of the $\mathbb{R}_j^2$ has an area of at least $2/3$, for otherwise the total volume would be less than $2/3 \cdot 1$. Thus, $(1/2, 1/2)$ is contained in the projection of b to $\mathbb{R}_j^2$, for all $1 \leq j \leq k$.

This means that every large box (of volume at least $2/3$) contains every point in all dimensions, except possibly for the corresponding $\mathbb{R}_i^2$. From this, the claim follows. □ □

Further, any box of volume $2/3$ has its lower left endpoint inside $[0, 1/2)^{2k}$. As all points lie inside $[1/2, 1]^{2k}$, we can extend any large empty until its "lower left" corner is the origin.

After these modifications, applying Lemma 6, we can use the same arguments as in Section 3: There is an empty box of volume C^k if and only if G has a k-clique. This proves the second part of Theorem 3.

6 The Box Discrepancy Problem

In order for our proof to work for this case, we will combine the ideas of the previous sections. Recall that we want to compute the box discrepancy

$$\bar{D}_{\mathcal{R}}(P) := \max_{R \in \mathcal{R}} \left| vol(R) - \frac{|R \cap P|}{|P|} \right|$$

of a point set P.

To simplify our arguments, we make sure that any box containing *all* points has volume 1. Thereto, we add one point at the origin and one point at $(1, 1, \ldots, 1)$.

Now we construct the point set with the constants determined in Section 4 (with N increasing by 2 because of the additional points). For this, we again choose C large enough so that the maximum is attained for a large box with *no points* inside, i.e., so that

$$C^k > \left(\frac{N-1}{N} \right).$$

Finally, we lift all points (except for the origin) as in Section 5. Using the same arguments, Theorem 2 follows.

7 Other geometric range spaces

So far we have considered range spaces determined by (restricted or unrestricted) boxes inside the unit cube, namely $\mathcal{B}$ and $\mathcal{B}_0$. Similar questions can be asked when the ranges are determined by other (geometric) objects. We give a few examples and a short discussion on how to adapt the proofs to these ranges. We restrict ourselves to the analog of the
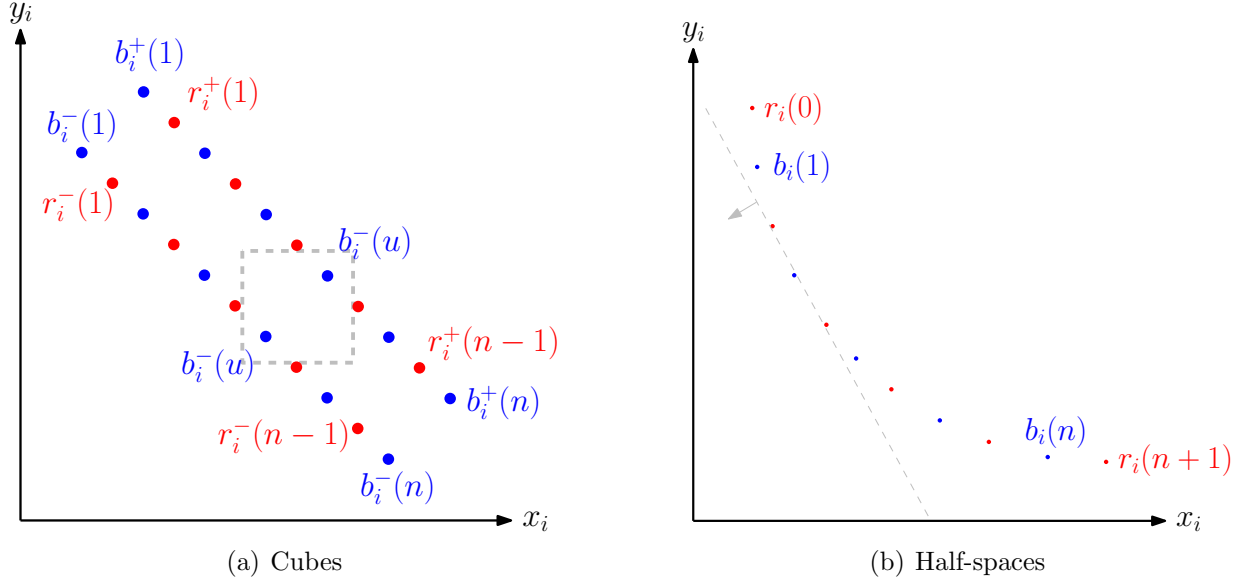
(a) Cubes   (b) Half-spaces

Figure 5: Other range spaces

BICHROMATIC-RECTANGLE problem, from which the hardness of computing the combinatorial discrepancy is easily derived. That means for a point set $P = P_r \uplus P_b$ and range space $\mathcal{R}$, we consider the value

$$E_{\mathcal{R}}(P_r, P_b) := \max_{R \in R, R \cap P_r = \emptyset} |R \cap P_b|.$$

Cubes If the range space is $\mathcal{Q}$, the set of all cubes inside the unit square, we can adapt the construction as shown in Figure 5(a).

Convex sets Here, the same arguments as in Section 2 work as well: Any convex set that does not contain any red points can contain at most one blue point from each $\mathbb{R}_i^2$. Further, also as a direct consequence of convexity, the encoding of the edges works as well.

Half-spaces Instead of putting all points on a line, we now put all points on a convex curve as in Figure 5(b). Note the two additional red points on both ends to prevent $b_i(1)$ and $b_i(n)$ to be chosen at the same time (by a hyperplane that does not contain the origin).

Theorem 7. *The problems d-CUBE-DISCREPANCY, d-CONVEX-SETS-DISCREPANCY and d-HALF-SPACE-DISCREPANCY are W[1]-hard with respect to d.*

8 Implications on verification of epsilon-nets

We concentrate on the problem d-NET-VERIFICATION for half-spaces. The proof can easily be adapted to the other range spaces considered.

Recall that, for given sets $S \subseteq P \subset \mathbb{R}^d$ and an $\varepsilon > 0$, we want to decide whether every half-space that contains at least $\varepsilon|P|$ points also contains a point from S.

To see why this problem is hard, consider the construction for the d-BICHROMATIC-RECTANGLE problem from Section 2. There, we had a set P_r of red and a set P_b of blue points. We have shown that it is W[1]-hard to decide whether there is a half-space containing k blue and no red points. To reduce this problem to d-NET-VERIFICATION, we set P to be the set of all points, and S to be the set of red points P_r. Then, we set $\varepsilon = k/|P|$. Now a half-space containing k points corresponds to a large set that is *not* intersected: it contains $k = \varepsilon|P|$ blue points but no red points. But this means that S is *not* an ε-net. This proves Theorem 5.

9   Conclusion

As the problems we have considered are all computationally hard when d is part of the input, we have to resort to approximation algorithms when dealing with them. We have seen that for the MAXIMUM-EMPTY-STAR problem, there is no hope for a polynomial time algorithm with a reasonable approximation factor at all. The (in)approximability of the (star) discrepancy is open even in the classical complexity theory framework. The question whether it can be approximated in FPT time when parameterized by dimension seems worthwhile considering.

References

[AHP08]   Boris Aronov and Sariel Har-Peled. On Approximating the Depth and Related Problems. *SIAM J. Comput.*, 38(3):899–921, 2008.

[Alo11]   Noga Alon. A Non-linear Lower Bound for Planar Epsilon-nets. *Discrete and Computational Geometry*, 2011. To appear. DOI:10.1007/s00454-010-9323-7.

[AS87]   Alok Aggarwal and Subhash Suri. Fast Algorithms for Computing the Largest Empty Rectangle. In *Symposium on Computational Geometry*, pages 278–290, 1987.

[BK09]   Jonathan Backer and J. Mark Keil. The Bichromatic Rectangle Problem in High Dimensions. In *Proc. of the 21st Annual Canadian Conference on Computational Geometry, Vancouver, BC, Canada*, pages 157–160, 2009.

[BK10]   Jonathan Backer and J. Mark Keil. The Mono- and Bichromatic Empty Rectangle and Square Problems in All Dimensions. In *Proc. of the 9th Latin American Theoretical Informatics Symposium*, volume 6034 of *LNCS*, pages 14–25, 2010.

[CCF+05]   Jianer Chen, Benny Chor, Mike Fellows, Xiuzhen Huang, David Juedes, Iyad A. Kanj, and Ge Xia. Tight lower bounds for certain parameterized NP-hard problems. *Inf. Comput.*, 201:216–231, 2005.

[CGK+11] Sergio Cabello, Panos Giannopoulos, Christian Knauer, Dániel Marx, and Günter Rote. Geometric clustering: fixed-parameter tractability and lower bounds with respect to the dimension. *ACM Transactions on Algorithms*, 2011. To appear.

[CGKR08] Sergio Cabello, Panos Giannopoulos, Christian Knauer, and Günter Rote. Geometric clustering: fixed-parameter tractability and lower bounds with respect to the dimension. In *Proc. 19th Ann. ACM-SIAM Sympos. Discrete Algorithms*, pages 836–843, 2008.

[Cha00] Bernard Chazelle. *The discrepancy method: randomness and complexity*. Cambridge University Press, 2000.

[CT97] Marco Cesati and Luca Trevisan. On the Efficiency of Polynomial Time Approximation Schemes. *Inf. Process. Lett.*, 64(4):165–171, 1997.

[DEM96] David P. Dobkin, David Eppstein, and Don P. Mitchell. Computing the Discrepancy with Applications to Supersampling Patterns. *ACM Trans. Graph.*, 15(4):354–376, 1996.

[DF99] Rod G. Downey and Michael R. Fellows. *Parameterized Complexity*. Monographs in Computer Science. Springer, 1999.

[DGW10] Benjamin Doerr, Michael Gnewuch, and Magnus Wahlström. Algorithmic construction of low-discrepancy point sets via dependent randomized rounding. *J. Complexity*, 26(5):490–507, 2010.

[DJ09] Adrian Dumitrescu and Minghui Jiang. On the largest empty axis-parallel box amidst n points. *CoRR*, abs/0909.3127, 2009.

[DT97] Michael Drmota and Robert F. Tichy. *Sequences, Discrepancies and Applications*, volume 1651 of *Lecture Notes in Mathematics*. Springer, 1997.

[EHL+02] Jonathan Eckstein, Peter L. Hammer, Ying Liu, Mikhail Nediak, and Bruno Simeone. The Maximum Box Problem and its Application to Data Analysis. *Comput. Optim. Appl.*, 23(3):285–298, 2002.

[FG06] Jörg Flum and Martin Grohe. *Parameterized Complexity Theory*. Springer, 2006.

[Gne08] Michael Gnewuch. Bracketing numbers for axis-parallel boxes and applications to geometric discrepancy. *J. Complexity*, 24(2):154–172, 2008.

[GSW09] Michael Gnewuch, Anand Srivastav, and Carola Winzen. Finding optimal volume subintervals with k points and calculating the star discrepancy are NP-hard problems. *J. Complexity*, 25:115–127, 2009.

[HW86]   David Haussler and Emo Welzl. Epsilon-nets and simplex range queries. In *Proceedings of the second annual symposium on Computational geometry*, SCG '86, pages 61–71. ACM, 1986.

[Mat10]   Jiří Matoušek. *Geometric Discrepancy: An Illustrated Guide*. Springer, 2010.

[Nie92]   Harald Niederreiter. *Random number generation and quasi-Monte Carlo methods*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[Nie06]   Rolf Niedermeier. *Invitation to Fixed-Parameter Algorithms*. Oxford University Press, 2006.

[PT10]   János Pach and Gábor Tardos. Tight lower bounds for the size of epsilon-nets. *CoRR*, abs/1012.1240, 2010.

[Thi01]   Eric Thiémard. An Algorithm to Compute Bounds for the Star Discrepancy. *J. Complexity*, 17(4):850–880, 2001.

[VC71]   Vladimir N. Vapnik and Alexey Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.